Motivations

• Internet of Things (IoT) is the expansion of Internet that impacts everyday lives

• It’s a network of Interconnected smart devices such as TV, Refrigerator, clock, and etc. which mostly are using Cloud storage as their storage medium.

• IoT use cases
  • Machine to machine interactions
  • Machine to human interactions

• Options to communicate with an IoT device
  • Graphical User Interface (GUI) which involves pushing the buttons or clicking
  • Speech interfaces
Main contributions

• Tools:
  • Speech recognition (acoustic models)
  • Natural Language Understanding (language models)

• Outcome:
  • Personalized Speech Recognition
    • By allowing user to customize their speech communications e.g. having names for devices
  • For smart home applications
  • And customizable devices
How Speech recognition works for IoT

Google Cloud Speech API
Alexa Voice Service (AVS)
Wit Speech API
CMUSphinx

Natural Language Processing
Statistical Language Models
Spoken Language Understanding for IoT

• Spoken Language Understanding (SLU) is the process of understanding human speech at machines
  1. Automatic Speech Recognition (ASR): Speech $\rightarrow$ Text
  2. Semantic Interpretation of what was said by the person

• Here, acoustic models and language models are used to decode speech signal into a sequence of words and then extract the intent from it
  • acoustic models extract sounds or phones (or diphones)
  • Language models captures linguistic units (or words)

https://cmusphinx.github.io/wiki/tutorialconcepts/
Language Models for IoT

• Types of language models
  • **Statistical Language Models (SLM)** which uses probability distribution of linguistic units
    - Large amount of training data
    - Flexible
  • **Context Free Grammars (CFG)** which uses linguistic rules (linguistic grammars)
    - End user interactions
    - Restrictive
  • **Domain-specific SLM** which uses a generic SLM plus domain specific knowledge
    - E.g. sequence of words that might use in a smart home
    - A same language model can be used for different users
    - Devices can be personalized by **device labels**
    - **Misrecognition of spoken language can happen**
  • **Dynamic Hierarchical Language Model (DHLM)**
    - Generic SLM + Domain knowledge + Device names

<table>
<thead>
<tr>
<th>Reference</th>
<th>Generic Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tell me status of the dinning room dimmer</td>
<td>Tell me status of the dinning room Denver</td>
</tr>
<tr>
<td>Turn gym light on</td>
<td>Turn Jim light on</td>
</tr>
<tr>
<td>Dim dinning room light</td>
<td>Jim dinning room light</td>
</tr>
<tr>
<td>What is the status of left kitchen window</td>
<td>What is the status of laugh kitchen window</td>
</tr>
</tbody>
</table>
Dynamic Hierarchical Language Model (DHLM)

- Hierarchical Language Model (HLM)
  - Creates a tree for the language model
    - Each level defines symbols undefined in a previous level
    - Each level can have undefined symbols
  - Statistical Language Model (SLM) is used to create the language model
  - (Weighted) Finite State Machine (FSM) is used to show language models
Dynamic Hierarchical Language Model (DHLM)

1. Top level SLM is a language model created for non-terminals (e.g. different devices are wildcarded)
2. Top level CFG is a CFG that covers common fixed phrases (e.g. help commands)
3. The subgrammars are CFGs that define the undefined symbols of the 1st level (e.g. replacing a wildcard in the language model with refrigerator)
Dynamic Hierarchical Language Model (DHLM)

• Different SLMs are build for different categories of devices such as lights, doors, windows, sensors, cameras, thermostats, and etc.
  • Each category has a set of various commands for training
• At train time, for each SLM the device category is replaced by the device name
• Requirements: Device names, User names

![Diagram of DHLM with nodes and edges labeled with commands like turn, dim, on, off, close, open, light, and door.]
Dynamic Hierarchical Language Model (DHLM)

• User database
  • User name
  • List of devices
    • Different variations of device names may be added to the model
    • List of categories of devices

• The SLM created from this information can be updated anytime

... Turn on _light
    Turn off _light
    Turn _light on
    Turn _light off
...

User 1

_userlight

Kitchen light
My favorite lamp
Master bedroom

User 2

_userlight

Jay’s room
Fish lamp
Semantic Analysis

• The output to the Spoken Language Understanding (SLU) is a semantic interpretation of what was said.
  • Speech recognition word accuracy is not so important
    • Why?
  • Semantic tags associated with the speech recognition output are more important
    • Why?
  • A successful task performs the correct action for the specific device
Semantic Analysis

• Semantic tags are used
  • Intent (the action part of the command)
  • Device name (the target device)
  • E.g. adjust the thermostat on the second floor to 68°
    • Intent: Set the temperature to 68° degrees
    • Device: Thermostat on the second floor

• How to calculate the accuracy
  • Compare the semantic tags VS human labeling
  • The annotated data (BRAT) is also used for training the top level SLM
Acoustic models

- Hidden Markov Model (HMM) to show the temporal variability of speech
  - And Gaussian Mixture Model (GMM) for each HMM state
  - GMM-HMM
- HMM + GMM + additional bottleneck features created by Deep Neural Network
  - DNN-GMM-HMM

\[
\text{Maximize } P(W|A) = P(A|W) \cdot P(W) / P(A)
\]

i.e. \( P(A|W) \) is the acoustic model e.g. HMM and \( P(A) \) is a constant
Results

• Each use connects to the system using a smartphone application
  • Each user has a set of devices and their names

• A cloud API for speech recognition receives the voice commands
  • A subset of the speech data is used to train the model

• Choice of acoustic models
  • GMM-HMM
  • DNN-GMM-HMM
Results

• Speech recognition **word accuracy**

<table>
<thead>
<tr>
<th>Language Model</th>
<th>Acoustic Model</th>
<th>Word Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DHLM</td>
<td>GMM-HMM</td>
<td>83.2</td>
</tr>
<tr>
<td>DHLM</td>
<td>DNN-GMM-HMM</td>
<td><strong>87.6</strong></td>
</tr>
<tr>
<td>Generic SLM</td>
<td>GMM-HMM</td>
<td>68.8</td>
</tr>
<tr>
<td>Generic SLM</td>
<td>DNN-GMM-HMM</td>
<td>81</td>
</tr>
</tbody>
</table>
Results

• **Semantic accuracy**
  - A task is accurate when **intent** and **device** are recognized correctly

• Incorrect intent, yet correct device
  - “turn on the kitchen light” → “turn off the kitchen light”
  - “set the temperature to 69°” → “set the temperature to 65°”

• Correct intent, yet incorrect device
  - “turn the **upstairs** thermostat off” → “turn the **downstairs** thermostat off”

• **Task is accurate** when both **intent** and **device** are recognized correctly

<table>
<thead>
<tr>
<th>Acoustic Model</th>
<th>Intent Accuracy (%)</th>
<th>Device Accuracy (%)</th>
<th>Task Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM-HMM</td>
<td>90.1</td>
<td>82.6</td>
<td>75.4</td>
</tr>
<tr>
<td>DNN-GMM-HMM</td>
<td>94.4</td>
<td>83.5</td>
<td>79.9</td>
</tr>
</tbody>
</table>
Conclusion

• Talking to an IoT device is an intuitive way to communicate with it
• Acoustic models and language models are used to bridge the gap between human and devices
• Personalized language models are necessary to have better accuracy in speech recognition process